Nota 5
Diferencias en Diferencias

1 Introducción

El método de diferencias en diferencias (Diff-in-Diff) se ha vuelto un método muy popular para hacer inferencia causal en microeconometría aplicada. El planteamiento de este método requiere observar dos grupos de individuos (o entidades) en al menos dos momentos distintos del tiempo, siendo uno de esos dos grupos afectado por un cambio, cuyo efecto causal se pretende estimar. Ejemplo: si el propósito fuera evaluar el efecto de una reforma o cambio legislativo utilizando el método de diferencias en diferencias, sería necesario observar a un grupo de individuos afectados por la reforma y a otro no afectado, antes y después del cambio. Existe también otra alternativa para aplicar el método de diferencias en diferencias sin tener observaciones en dos momentos del tiempo. Ésta consiste en tomar dos subgrupos de individuos tanto en el grupo afectado y el no afectado por el cambio: individuos elegibles y no elegibles para ser afectados por el cambio. Ejemplo: supongamos que se implementa un programa de construcción de escuelas en un Estado que beneficia a niños indigenas. Podríamos aplicar el método de diferencias en diferencias si tenemos una base de datos con observaciones de niños indígenas y no indígenas en el Estado donde se aplico la reforma y en al menos otro Estado no afectado.

Es común pensar en aplicar el método de diferencias en diferencias cuando se tiene un *experimento natural*. Un *experimento natural* consiste en un evento exógeno que afecta ciertas variables económicas (en términos econométricos, alguna variable independiente). La exogeneidad del evento consiste en que éste debe afectar a la variable independiente en cuestión y no a otros variables independientes que pudieran afectar la variable dependiente que se analiza. Ejemplos de experimentos naturales: eventos naturales como sequías, proliferación de algún virus, desastres naturales, eventos históricos etc.

2 Planteamiento básico

Empezando por el caso más simple, supongamos que observamos a individuos que se dividen en dos grupos: $G_i = \{0, 1\}$, donde el grupo 1 es afectado por el cambio y el 0 no; y que pueden ser observados en dos momentos del tiempo $T_i = \{0, 1\}$, donde $T_i = 0$ es antes del cambio y $T_i = 1$ después. (Nota: No es necesario que la base de datos sea de panel, es decir, que sea el mismo individuo el que sea observado en ambos momentos del tiempo. Únicamente se requiere el supuesto de que la muestra sea aleatoria) Para dichos individuos observamos una variable de interés: Y_i. Por lo tanto, cada observación estará caracterizada por tres variables (Y_i, G_i, T_i). Supongamos que queremos ver el impacto de una reforma laboral sobre horas trabajadas y suponemos que dicha reforma se puede aplicar en unos Estados si $(G_i = 1)$ y en otros no $(G_i = 0)$. Un individuo que trabaja 20 horas a la semana, observado antes de la reforma en un Estado donde si se aplica la reforma, tendrá las variables: $(20, 1, 0)$.

Para calcular el efecto de la reforma laboral utilizando el modelo de diferencias en diferencias primero tendremos que calcular las medias de cada grupo en ambos momentos del tiempo. Es decir, generaremos el siguiente estadístico:

$$\overline{Y}(j,t) = \frac{\sum_{i=1}^{N} 1(G_i=j)*1(T_i=t)*Y_i}{\sum_{i=1}^{N} 1(G_i=j)*1(T_i=t)}, \text{ para } j = \{0,1\} \text{ y } t = \{0,1\}$$

En particular, $\overline{Y}(0,1)$ es la media de horas trabajadas de los individuos observados después de la reforma en el Estado B.

El estimador de diferencias en diferencias es:

$$\tau = [\overline{Y}(1,1) - \overline{Y}(1,0)] - [\overline{Y}(0,1) - \overline{Y}(0,0)]$$

Intuitivamente, τ representa la diferencia del cambio promedio de horas trabajadas en el estado afectado respecto al no afectado. Como vimos en la *Nota 2*, este mismo efecto lo podemos estimar con una regresión de MCO si utilizamos la siguiente especificación:

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 T_i + \tau G_i T_i + U_i \tag{1}$$

En este caso, si τ es significativo, podremos decir que la reforma laboral tuvo un efecto estadísticamente significativo en las horas trabajadas.

3 Supuesto de tendencia paralela

En el planteamiento del modelo, el grupo no afectado por el cambio (el grupo $G_i = 0$ en nuestro ejemplo) funciona como un "grupo de control." Si no tuviéramos la información del grupo de control, nuestra única alternativa consistiría en comparar la variable dependiente Y_i antes y después del cambio en el grupo afectado. Sin embargo, en este caso, no sería posible distinguir si el cambio en el nivel medio de Y_i se debió a un cambio económico a través del tiempo o a el cambio específico que estamos analizando.

En el ejemplo de la reforma laboral, si no contaramos con el grupo de control y quisiéramos estimar el efecto de la reforma tomado el cambio en el promedio de horas trabajadas antes y después de la reforma laboral en el grupo $G_i = 1$, calcularíamos $[\overline{Y}(1,1) - \overline{Y}(1,0)]$. Sin embargo, utilizando este estimador sería imposible distinguir si el cambio en horas trabajadas se debe a la reforma laboral o a cualquier otro cambio sucedido entre el año 0 y 1 (e.g. una desaceleración económica, elecciones políticas, etc.).

Nuestro grupo de control nos permite controlar precisamente por esos efectos. Para ello, nuestro supuesto clave es que "en ausencia del cambio (reforma laboral), el grupo afectado por el cambio ($G_i = 1$), hubiera tenido una tendencia igual a la que tuvo el grupo de control ($G_i = 0$)". Este supuesto se conoce como el **supuesto de tendencia paralela**. En términos de nuestras ecuaciones, este supuesto genera un *contrafactual*. Es decir, asume que si no hubiese habido reforma, el cambio de horas trabajadas en el grupo $G_i = 1$ hubiera sido $[\overline{Y}(0,1) - \overline{Y}(0,0)]$. Como si hubo reforma, el cambio observado fue $[\overline{Y}(1,1) - \overline{Y}(1,0)]$. Por lo tanto, el efecto de la reforma es la diferencia entre dichos valores, es decir, τ.

Para que el supuesto de tendencias paralelas sea válido el grupo de control y el afectado deben ser lo más parecidos posibles antes del cambio. En función de nuestra especificación [1] esto debería querer decir que preferentemente β_1 debe ser no significativo. Otra manera de comprobar esto es hacer una comparación de medias con otras variables observables entre ambos grupos. Esta es una prueba sencilla y consiste en calcular los valores medios de distintas variables disponibles en nuestra base de datos antes del cambio (i.e. cuando $T_i = 0$). Si ambos grupos son similares antes del cambio o la reforma, la diferencia de medias no debe ser estadísticamente significativo o distinto a cero en la mayoría de las variables.

Otra posibilidad para hacer el calculo del efecto de diferencias en diferencias consiste en utilizar el modelo de *efectos fijos*. Supongamos que se tiene información para todos los estados j de algún país y que la reforma se aplicó en un subconjunto de esos estados. Sea R_{jt} una dummy que indica si en el año t y el estado j, la ley ya se encontraba vigente. Para estimar el efecto de la ley utilizando el modelo de efectos fijos únicamente estimamos la siguiente especificación utilizando MCO:

$$Y_{ijt} = \alpha_j + \delta_t + \tau R_{jt} + X'_{ijt}\beta + U_{ijt} \tag{2}$$

En este caso, los subíndices i corresponden al individuo, j al estado y t al tiempo; α_j es un efecto fijo por estado e indica que se debe incluir una dummy por estado; δ_t indica similarmente que se debe incluir una dummy por año para controlar por el efecto de tiempo; $X'_{ijt}\beta$ son controles a nivel individuo que cambian de estado a estado y a lo largo del tiempo; U_{ijt} es el error que en este caso debe incluir *cluster* a nivel estado; y τ es el efecto de diferencias en diferencias que nos interesa estimar.

El modelo de efectos fijos además permite estimar el efecto de algún cambio cuando este se da en distintas intensidades en diferentes grupos. Por ejemplo, supongamos que queremos estimar el efecto de un impuesto al consumo de alcohol que se empezó a establecer en los estados, pero cada estado lo implementó con diferente intensidad. En la especificación anterior, esto querría decir que ahora R_{jt} es simplemente el nivel del impuesto en el estado j y año t; Y_{ijt} puede ser consumo de alcohol; y el resto de las variables se describe igual.

4 Pruebas de robustez

Para verificar qué tan robustos son los resultados del método de diferencias en diferencias es común (y recomendable) llevar a cabo las siguientes pruebas:

1. Utilizar datos de más de un periodo previo al cambio y llevar a cabo un ejercicio de diferencias en diferencias. Con esta prueba se puede evaluar si la tendencia de ambos grupos antes del cambio sigue una tendencia paralela. Si se cuentan con varios periodos de información, puede utilizarse la siguiente especificación:

$$Y_{ijt} = \alpha_j + \delta_t + \sum_{t=1}^{T} \tau_t T_t R_j + X'_{ijt}\beta + U_{ijt}$$

   donde T_t es una dummy igual a uno si la observación se lleva a cabo en el periodo t y R_j es una dummy igual a uno si el estado j es del grupo de estados que implementaron la reforma. En este caso, si el cambio tuvo un impacto significativo y se dio en el año $t' \in (1, T)$, deberían observarse τ_s no significativas si $s < t'$ y τ_r significativas si $r \geq t'$.

2. Utilizar un grupo de control alternativo. Si los resultados son distintos utilizando un grupo de control u otro, esto puede representar un problema de validez. En clase ejemplificaremos esto utilizando triples diferencias. La idea de triples diferencias (DDD) es comparar dos estimadores de diferencias en diferencias (DD).

3. Hacer el mismo ejercicio de diferencias en diferencias utilizando una variable dependiente que no debería haber sido afectada por el cambio. Esto se conoce como *prueba de falsificación*.

5 Problemas comunes de diferencias en diferencias

Los críticas más usuales a tener en cuenta con el modelo de diferencias en diferencias son:

1. Que el cambio (o reforma) se produzca como resultado de condiciones pre-existentes.

   (a) Targeting basado en las condiciones preexistentes. Supongamos que queremos ver el efecto de una reforma educativa sobre la cobertura escolar. Esta reforma

consiste en construir escuelas y para llevarla a cabo se eligió a los 5 estados de la república. El modelo de diferencias en diferencias consistirá en comparar la cobertura escolar en los estados elegidos con los no elegidos antes y después de la implementación del programa. Sin embargo, si la selección de los estados para la reforma se llevó a cabo por que en dichos estados había mas presión por parte de los padres de familia, habría un problema de identificación. No será posible distinguir si fue la exigencia y preocupación de los padres, características de dichos estados que impulsaron la exigencia de los padres o la propia construcción de escuelas lo que generó los resultados del modelo de diferencias en diferencias.

(b) *Ashenfelter dip.* Este efecto toma su nombre de trabajos de investigación de Orley Ashenfelter y David Card que analizaron el efecto de programas de entrenamiento sobre el salario. Los investigadores se dieron cuenta que los participantes del programa habían decidido inscribirse a los programas de entrenamiento ya que su salario había disminuido recientemente. Por lo tanto, al estimar un modelo de diferencias en diferencias el efecto se incrementaba debido a que el grupo que participaba en el entrenamiento tenía salarios que habían disminuido recientemente antes del programa.

2. Uso de la forma funcional. Tomaremos un ejemplo en el cual después de una intervención el desempleo bajo de 30% a 20% en el grupo que fue afectado por una reforma, mientras que el desempleo en el grupo de control paso de 10% a 5%. Compararemos que pasa si utilizamos los valores medidos en tasa o si utilizamos logaritmos.

3. Comparación de efectos en corto versus largo plazo. En muchos casos, la pregunta relevante de alguna política es interesante respecto a su efecto de largo o mediano plazo. Sin embargo, el supuesto de tendencia paralela es mas realista y fácil de justificar en el corto plazo.

4. Efectos heterogeneos entre ambos grupos. El modelo de diferencias en diferencias puede también aplicarse si los dos grupos fueron afectados por un cambio pero de distintas magnitudes (como se explico en el ejemplo del impuesto al consumo de alcohol en la sección 2). En este caso, podría existir un problema si los efectos del cambio son heterogeneos entre los grupos. Este es un caso particular de una violación

al supuesto de tendencia paralela.